ENNAACT is a novel tool which employs neural networks for anticancer activity classification for therapeutic peptides

Patrick Brendan Timmons, Chandralal M. Hewage*

UCD School of Biomolecular and Biomedical Science, UCD Centre for Synthesis and Chemical Biology, UCD Conway Institute, University College Dublin, Dublin 4, Ireland

ARTICLEINFO

Keywords:
Neural network
Machine learning
Anticancer drugs
Peptides
In silico screening

ABSTRACT

The prevalence of cancer as a threat to human life, responsible for 9.6 million deaths worldwide in 2018, motivates the search for new anticancer agents. While many options are currently available for treatment, these are often expensive and impact the human body unfavourably. Anticancer peptides represent a promising emerging field of anticancer therapeutics, which are characterized by favourable toxicity profile. The development of accurate *in silico* methods for anticancer peptide prediction is of paramount importance, as the amount of available sequence data is growing each year. This study leverages advances in machine learning research to produce a novel sequence-based deep neural network classifier for anticancer peptide activity. The classifier achieves performance comparable to the best-in-class, with a cross-validated accuracy of 98.3%, Matthews correlation coefficient of 0.91 and an Area Under the Curve of 0.95. This innovative classifier is available as a web server at https://research.timmons.eu/ennaact, facilitating *in silico* screening and design of new anticancer peptide chemotherapeutics by the research community.

1. Introduction

Cancer is a leading cause of death worldwide, second only to cardiovascular diseases [1,2], and accounted for 9.6 million deaths in 2018. While traditional anticancer therapies, such as chemotherapy and radiation, are highly efficacious, they are nonetheless expensive and exert undesirable side-effects on non-cancerous cells. Additionally, cancer cells can develop resistance towards conventional chemotherapeutics [3]. The development of novel anticancer agents is, therefore, a goal of great importance [4,5].

Host defence peptides (HDPs) form an ancient element of the immune response of all life forms, having been identified in all species [6, 7]. Indeed, the peptides' inductions pathways are highly conserved among the genomes of plant, insect and vertebrate species [7]. These peptides are a promising class of anticancer agents as they possess a number of advantages over non-peptide drugs, namely high specificity, cost-effective production, high tumour penetration, ease of modification and synthesis, and an insusceptibility to cancer drug resistance [8,9]. Anticancer peptides (ACPs) are usually short, typically 5–30 amino acids in length, and are often derived from HDPs. Peptides with anticancer activity have been found to possess α-helical, β-sheet and extended conformations in the presence of membranes or membranomimetic environments have been observed [10–13]. While they are typically cationic, anionic ACPs have also been found in nature [14,15]. Studies have previously shown that many HDPs which possess antimicrobial activity also exhibit selective toxicity towards cancer cells [13].

The mechanism by which ACPs exert their anticancer activities is a controversial topic which is not yet fully understood. A number of prospective mechanisms of action have been proposed to explain the peptides' oncolytic effects, which can be classed as either membranolytic or non-membranolytic. These mechanisms include cytoplasmic membrane disruption via micellization or pore formation, and the induction of apoptosis via disruption of the mitochondrial membrane [9,16]. The precise nature of the mechanism used depends on the peptide's physicochemical properties and the characteristic of the target membrane [17,18]. Indeed, the different physicochemical properties of malignant and healthy mammalian cells are responsible for the peptides' observed selectivity towards cancer cells [18]. Unlike healthy cells, which possess a zwitterionic cell membrane, cancerous cells are characterized by a cell membrane which possesses a net negative charge, which is conferred to it by anionic molecules such as phosphatidylserine, O-glycosylated mucins, sialylated gangliosides and heparin sulfate [13, 18,19]. Cancer cells' relatively high cell-surface area and increased membrane fluidity result in the binding of a greater number of peptides and enhanced lytic activity [20,21].

In silico methods facilitate an efficient and accelerated avenue for exploring the large chemical space that peptides occupy, by reducing the number of peptides that need to be experimentally screened for anticancer activity. Previously, multiple methods for the prediction of a peptide's anticancer activity have been proposed and described in the literature. These methods include AntiCP [22], AntiCP 2.0 [23], iACP [24], ACPP [25], iACP-GAEnsC [26], MLACP [27], SAP [28], TargetACP [29], ACPred [30], ACPred-Fuse [31], ACPred-FL [32], PTPD [33], PePred-Suite [34], mACPpred [35], ACP-DL [36], the methods of Li and Wang [37], Hajisharifi et al. [38], Khan et al. [39], Gabernet et al. [40] and Li et al. [41]. The application of machine learning to the study of anticancer peptide bioactivity has been comprehensively reviewed by Shoombuatong et al. [42]. Most use similar dataset construction methods and employ support vector machines (SVM) or random forests (RF) as the machine learning classifier. A variety of different feature

* Corresponding author.
E-mail address: chandralal.hewage@ucd.ie (C.M. Hewage).

https://doi.org/10.1016/j.biopha.2020.111051
Received 28 August 2020; Received in revised form 8 October 2020; Accepted 19 November 2020
Available online 27 November 2020
0753-3322/© 2020 The Authors. Published by Elsevier Masson SAS. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

representations have been employed in these studies, including physicochemical features, amino acid composition, dipeptide composition, and Chou's pseudo amino acid composition [43]. Many other areas of bioinformatics have benefited from the predictive power of deep learning, with neural network classifiers existing for multiple tasks, such as DeepPPISP for the prediction of protein–protein interaction sites [44], SCLpred for protein subcellular localization prediction [45] and CPPpred for the prediction of cell-penetrating peptides [46]. Furthermore, this work complements our existing machine learning studies of hemolytic activity in antimicrobial peptides and builds on the neural network architecture employed in the HAPPENN classifier [47]. With the amount of anticancer sequence data ever-increasing, we have exploited the available data to create a deep neural network for classifying peptides as anticancer or non-anticancer on the basis of the primary sequence. Herein, we present ENNAACT, a novel neural network peptide anticancer activity classifier, which is available for the benefit of the academic community as a free-to-use online web server at https://re search.timmons.eu/ennaact.

2. Methods

The development of a machine learning classifier should satisfy five criteria, collectively known as Chou's 5-steps rule [48]: (i) a valid dataset must be constructed to train and validate the predictor, (ii) an effective representation should be developed that possesses a relationship to the target to be predicted, (iii) a machine learning model should be trained on the representation of the dataset, (iv) the predictive performance should be evaluated by means of cross-validation, (v) the predictor should be made available as a web server. The methods employed in creating ENNAACT satisfy Chou's 5-steps rule.

2.1. Datasets

The construction of a reliable dataset is an important step in a machine learning endeavour. Neural networks thrive on large datasets, it was therefore imperative to construct as a dataset encompassing as many anticancer peptides as possible.

Peptide sequences in the dataset consist only of natural amino acids, with only the N-terminal acetylation and C-terminal amidation modifications considered. Information about the peptides' secondary structure is not included in the dataset. The dataset is available for download from the webserver website and as supplementary material to this article.

Anticancer peptides were extracted from the DBAASP [49], dbAMP [50], CancerPPD [51] and BioPepDB [52] databases. As no resource exists that details peptides with an experimentally confirmed absence of anticancer activity, random non-secretory sequences from the Universal Protein Resource were used as negative dataset, as anticancer peptides are usually secretory in nature [53].

Additionally, both datasets were supplemented with peptides from the datasets of Hajisharifi et al. [38], Basith et al. [54], ACPred-FL [32], ACPred [30], iACP [24], mACPpred [35], AntiCP [22] and AntiCP 2.0 [23].

To avoid bias in the classifier, the dataset was internally redundancy reduced using CD-HIT [55–57], with a sequence identity cut-off of 0.9. The final ENNAACT dataset consists of 659 experimentally validated anticancer peptides and 5298 non-anticancer peptides between 7-40 amino acids in length, and their corresponding anticancer/non-anticancer designations.

2.1.1. Datasets for additional benchmarking

ENNAACT-HARD dataset. Compositionally similar peptides with differing biological activities pose a significant challenge for the development of machine learning-based predictors. In order to ascertain the model's capacity to discriminate between such similar peptides, an additional dataset was created, ENNAACT-HARD, which consists of the 659 experimentally validated anticancer peptides from the ENNAACT dataset, and ENNAACT's non-anticancer peptides with the greatest compositional similarity to the anticancer peptides, as measured by their Euclidean distances to the anticancer sequences [58–60,47].

ENNAACT+A1. The recently published AntiCP 2.0 is considered to be representative of the best-in-class of anticancer activity predictors. In order to fairly compare the performance of the ENNAACT predictive method to the results, an additional dataset was constructed which consists of the AntiCP 2.0 test dataset's peptides as the independent test set. The training dataset consists of the peptides sequences of the redundancy-reduced ENNAACT dataset, which possess < 0.9 similarity to the AntiCP 2.0 test peptides, as calculated by the CD-HIT-2D program resulting in a set of 574 anticancer sequences and 5176 non-anticancer sequences. The AntiCP 2.0-derived test set consists of 181 anticancer peptides and 134 non-anticancer peptide sequences containing exclusively proteinogenic amino acids, and between 7 and 40 amino acids in length.

ENNAACT+A2. In order to establish the fairest comparison possible to the AntiCP 2.0 method, a dataset containing only the peptide sequences of the AntiCP 2.0 dataset was created. The sequences included contain 7–40 exclusively proteinogenic amino acids. This dataset contains 716 anticancer and 571 non-anticancer sequences for training, and 181 anticancer and 134 non-anticancer sequences for external validation. To maintain comparability to AntiCP 2.0, this dataset is not redundancy reduced.

ENNAACT+B. A study was recently performed by Basith et al. [54] which used an independent benchmark dataset to assess the performance of existing anticancer peptide prediction methods. To compare the performance of the ENNAACT predictive method to the results of that study, an additional dataset was created that comprises the Basith benchmark dataset as the independent test set. The training dataset comprises the sequences of the ENNAACT dataset which possess <0.9 and <0.6 similarity to the Basith benchmark dataset positive and negative samples, respectively, as

calculated by the CD-HIT-2D program, resulting in a set of 507 anticancer sequences and 3623 non-anticancer sequences. The test set consists of 236 anticancer peptides and 1733 non-anticancer peptide sequences containing 7–40 exclusively proteinogenic amino acids.

2.2. Model validation

Thorough validation of classifier models created by machine learning is of utmost importance. Consequently, 10-fold cross- validations and validation by an external test set were employed for the performance evaluation of all models presented herein. The ENNAACT dataset was split into twelve parts, ten of which were used for cross-validation, whereby one of the subsets was validated upon while the other nine subsets were trained upon. The models trained under cross-validation were ensembled and evaluated with the external test set, which consists of the previously unused two of the twelve subsets.

2.3. Amino acid composition analysis

The amino acid composition of the anticancer peptides was analysed and compared to that of the non-anticancer peptide sequences. The analysis comprises the peptides' full sequences, the 15 N-terminal residues, and the C-terminal 15 residues.

2.4. Residue position preference analysis

To identify position-specific amino acid preferences, enrichment depletion logos (EDLogo) [61] were created for the anticancer peptides' sequences. The non-anticancer peptide sequences were used as the baseline in the logo plots' construction.

2.5. Features extraction

A selection of features was extracted from the peptides' primary sequences, which can be divided into two subcategories, physicochemical descriptors and amino acid-based descriptors. Only features which were non-zero for at least 50 samples were retained in the final feature vector.

2.5.1. Physicochemical descriptors

Global physicochemical descriptors,as well as amino acid scale- based descriptors, were calculated using the modlAMP package [62]. The former includes features such as molecular formula, sequence length, molecular weight, instability index, aliphatic index [63], aromaticity index [64], hydrophobic ratio, Boman index [65], sequence charge, charge density and isoelectric point. The latter, meanwhile, include hydrophobicity [66–70], side-chain bulkiness [71], pepArc [62], refractivity [72], side-chain flexibility [73], α-helix propensity [74], transmembrane propensity [75], polarity [76,71], amino acid charges, AASI [77], ABHPRK [62], COUGAR [62], Ez [78], ISAECI [79], MSS [80], MSW [81], PPCALI [82], t_scale [83], z3 [84] and z5 [85].

Additional physicochemical features were calculated based on amino acid properties detailed in the AAindex [86].

Sequence hydrophilicity was characterized using properties based on the amino acid hydrophilicity scale [87], the amino acids' net charges [88], polar requirements [89] and fractions of site occupied by water [90].

Similarly, sequence hydrophobicities were quantified using the amino acids' hydrophobicities [91,92], hydropathies [93], retention coefficients in HPLC [94] and partition energies [95,96].

Steric-related descriptors were calculated based on the residues' steric hindrance [97] and bulkiness [71], while secondary structure features were calculated based on helical [98] propensities.

As membrane interaction generally plays an important role in peptides' mechanisms of action, features based on membrane-buried preference parameters [99] and the side-chain interaction parameters [100] were calculated.

2.5.2. Composition descriptors

Compositional descriptors were calculated based on the amino acid, dipeptide, and tripeptide compositions for the conventional 20-amino acid alphabet, as well as the reduced alphabets of Thomas and Dill [101], Veltri et al. [102], and the conjoint alphabet [103]. *g*-Gap dipeptide compositions were calculated to account for the three-dimensional structure of the peptides [104], with the values of the parameter *g* being 1, 2 and 3. Furthermore, conjoint triad, composition, transition and distribution [105] and pseudo amino acid composition [106] descriptors were also calculated.

2.6. Machine learning approaches

Multiple machine learning approaches are employed in the current study, specifically principal component analysis (PCA) [107], t-Distributed Stochastic Neighbour Embedding (t-SNE) [108] to visualize the data in an unsupervised manner, and support vector machine (SVM) [109], random forest (RF) [110], and dense fully connected neural networks [111] for creating supervised classifiers.

SVMs were trialled using both a linear and non-linear radial base function (RBF) kernel. A grid search was employed for the tuning of the SVM regularization parameter C and kernel width parameter γ. Random forest hyperparameters, namely number of estimators, the maximum number of features, and the maximum depth were also adjusted using a grid search strategy.

2.7. Neural network architecture and implementation

Keras, a popular deep learning framework, was used to build the deep, fully-connected neural network. Tensorflow [112] was used as the back-end.

The neural network's input features are scaled to have minimum and maximum values of 0 and 1, respectively.

A randomized grid search was used for the identification of the optimal neural network architecture and hyperparameters. The first and final hidden layers have 64 nodes and are separated by six hidden layers of 256 nodes each. Batch normalization [113] is applied before the ReLU activation function for each hidden layer, which in each case is followed by a Dropout layer, with a rate of 0.80, which serves to prevent overfitting [114]. The output layer is a single node with a sigmoid activation function.

This binary cross-entropy loss function is frequently employed in binary classification neural networks. As the predicted labels of all training data approach their respective true values, the value of the function approaches zero. For this reason, the binary cross-entropy loss function was applied to this neural network. It is defined as:

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i log(y_i) + (1-y_i) log(1-\hat{y}_i)]$$

where y_i is the true value of the ith sample, and $\hat{y}_i$ is the predicted value of the ith sample. The loss function was weighted to adjust for the imbalanced number of positive and negative samples during training.

Adaptive Momentum (Adam) was identified as the optimal optimizer. Adam utilizes the following formula to update the neural network weights [115]:

$$\Theta_{t+1} = \Theta_t - \frac{\eta \hat{m}^t}{\sqrt{\hat{v}_t} + \epsilon}$$ where the $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected estimates of the mean and the variance of the gradients, respectively. The

The neural network was trained for 300 epochs, without stopping criteria. The model with the highest validation accuracy encountered during training was retained for each of the cross-validation splits. The optimal initial learning rate was found to be 0.1, with a decay of 0.0001.

2.8. Performance evaluation

The classifier's robustness is evaluated by a number of standard metrics, including metrics capable of evaluating imbalanced datasets, namely accuracy (Acc), sensitivity (Sn), specificity (Sp), the Matthews correlation coefficient (MCC), and by the receiver operating characteristic (ROC) curve. Confidence intervals are provided at the 95% level of significance.

The first four metrics are defined by the following equations:

$$\text{Acc} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100\% \quad (3)$$

$$\text{Sn} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\% \quad (4)$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100\% \quad (5)$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (6)$$

where

- TP = True positives: the number of correctly predicted positive (anticancer) peptides.
- FP = False positives: the number of non-anticancer peptides incorrectly predicted as being anticancer.
- TN = True negatives: the number of correctly predicted negative (non-anticancer) peptides.
- FN = False negatives: the number of anticancer peptides incorrectly predicted as being non-anticancer.

3. Results

The ENNAACT dataset was constructed from peptide sequences whose anticancer activity has previously been experimentally validated, and complemented by random, non-secretory sequences extracted from the Universal Protein Resource which were used as the negative examples.

Illustrating the relationship between dataset's peptides' features and their corresponding anticancer activity facilitates analysis of this complicated relationship [116]. The peptide sequence dataset was subjected to an amino acid composition analysis (Fig. 1) and residue position preference analysis (Fig. 2). Feature vectors were constructed for the peptides based on the sequences' physicochemical and compositional descriptors. Principal component analysis (PCA) (Fig. 3) and t-stochastic neighbour embedding (t-SNE) (Fig. 4) plots are used to visualize the feature vectors in two-dimensional space, both of which show an incomplete separation of the positive and negative classes. Finally, three machine learning classifiers, namely support vector machines, random forests and neural networks, are trained on the dataset's feature vectors, and the anticancer activity prediction results are evaluated.

3.1. Amino acid composition analysis

In order to determine if certain residues are preferred in anticancer peptides sequences compared to non-anticancer sequences, an amino acid residue composition analysis was carried out; its results are illustrated in Fig. 1. Statistical analysis was carried out on the results using an $N-1$ Chi-squared test; all results are significant at the $P < 0.05$ significance level.

Most interestingly, anticancer peptides are most enriched in the positively charged

lysine residue, but are only enriched in arginine to a much smaller extent. Anticancer peptides are also enriched in the hydrophobic leucine, isoleucine, phenylalanine and tryptophan residues, but are depleted in alanine, valine, methionine and tyrosine.

Significant enrichment is also observed in cysteine. The greatest depletion is observed in the negatively charged aspartic and glutamic acid residues, with depletion also seen in the histidine, asparagine, glutamine, proline, serine and threonine. Anticancer peptides are overall slightly enriched in glycine.

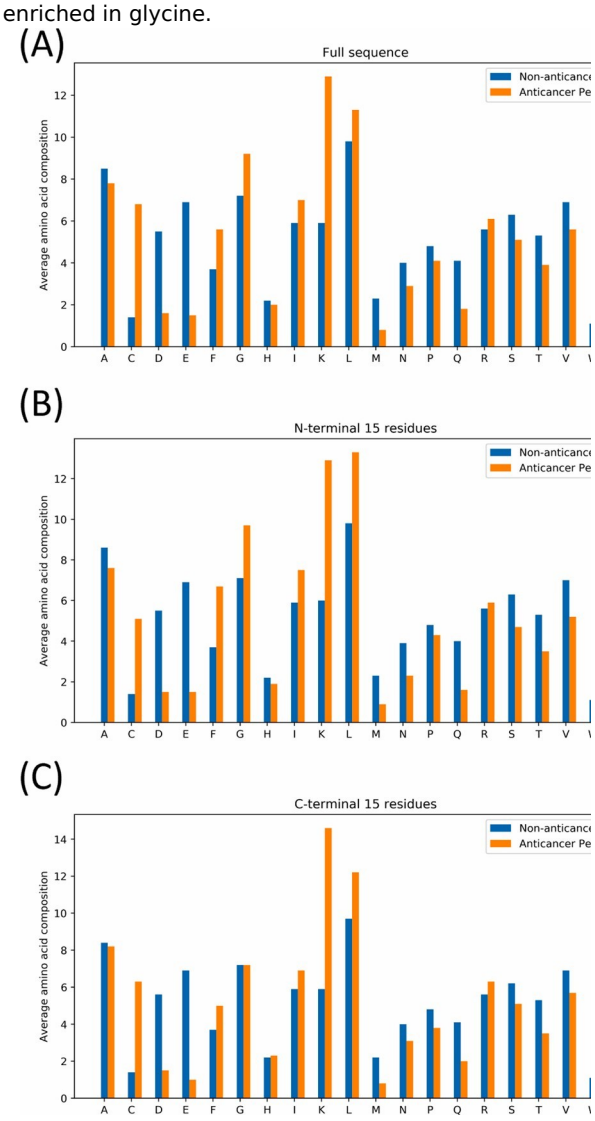
(A)
Full sequence
Non-anticanc
Anticancer Pe
Average amino acid composition
A C D E F G H I K L M N P Q R S T V
(B)
N-terminal 15 residues
Non-anticanc
Anticancer Pe
Average amino acid composition
A C D E F G H I K L M N P Q R S T V
(C)
C-terminal 15 residues
Non-anticanc
Anticancer Pe
Average amino acid composition
A C D E F G H I K L M N P Q R S T V

Fig. 1. Percentage average amino acid residue composition of the (A) full sequences, (B) N-terminal 15 residues, and (C) C-terminal 15 residues of experimentally validated anticancer peptides (orange), and non-anticancer peptide sequences randomly extracted from UniProt proteins (blue). One-letter amino acid codes are given for the residues on the *x*-axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Residue position preference analysis

To determine if a preference exists for certain residues at certain positions in the peptide sequences, an enrichment-depletion logo plot was created (Fig. 2) for the anticancer peptide sequences, with the non- anticancer class serving as the baseline for the plot. Enriched residues are those which are more common at that position in the anticancer sequences, relative to the non-anticancer class, and depleted residues are those which are less common.

Initial inspection of the results reveals information consistent with that gleaned from the amino acid composition analysis: anticancer peptides are enriched in the positively charged lysine residue, cysteine, and in the hydrophobic tryptophan and isoleucine, residues, while being depleted in the negatively charged residues, with the single exception of a slight enrichment in aspartic acid at position 4.

Further inspection leads to the appearance of more position-specific enrichments. A preference exists at position 1 for glycine and phenylalanine, positions 2 and 3 for tryptophan, positions 3 and 14 for proline,

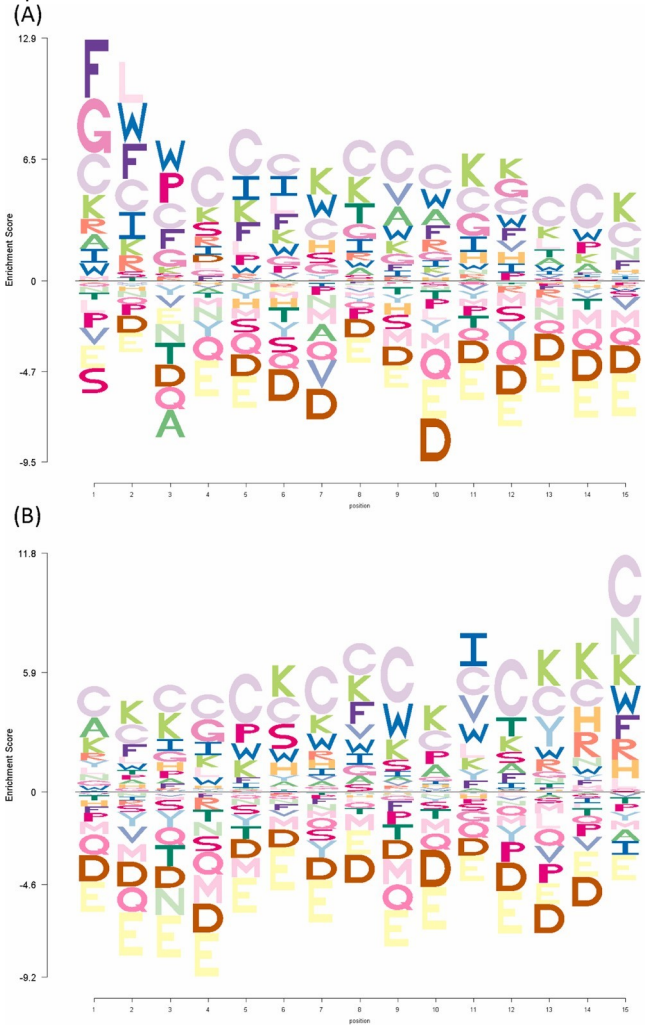

Fig. 2. Enrichment-depletion logo plot of (A) N-terminal 15 residues and (B) C- terminal 15 residues of experimentally validated anticancer peptides of the ENNAACT dataset. Data is scaled to account for the background probability of each amino acid, based on the negative dataset.

position 4 and 7 for serine, position 8 for threonine and position 9 for valine. A notable preference also exists for asparagine at the C-terminal position,

histidine at the second-last position and tyrosine at the third- last position.

3.3. Data visualization

3.3.1. Principal component analysis (PCA)

A principal component analysis (PCA) was carried out for the set of all computed descriptors, as well as the physicochemical and compositional descriptor subsets (Fig. 3). Inspection of the PCA plots suggests that a separation does exist between the anticancer and non-anticancer classes, but the separation is not complete and a significant overlap exists between the two classes.

3.3.2. t-Distributed Stochastic Neighbour Embedding (t-SNE) In addition to the PCA analysis, a t-Distributed Stochastic Neighbour Embedding (t-SNE) analysis was carried out for the set of all computed descriptors, as well as the physicochemical and compositional descriptor subsets (Fig. 4). Similarly to the PCA results, an incomplete, albeit larger, separation exists between the anticancer and non-anticancer classes in all three datasets. The best separation between the two classes appears to be in the physicochemical descriptor subset.

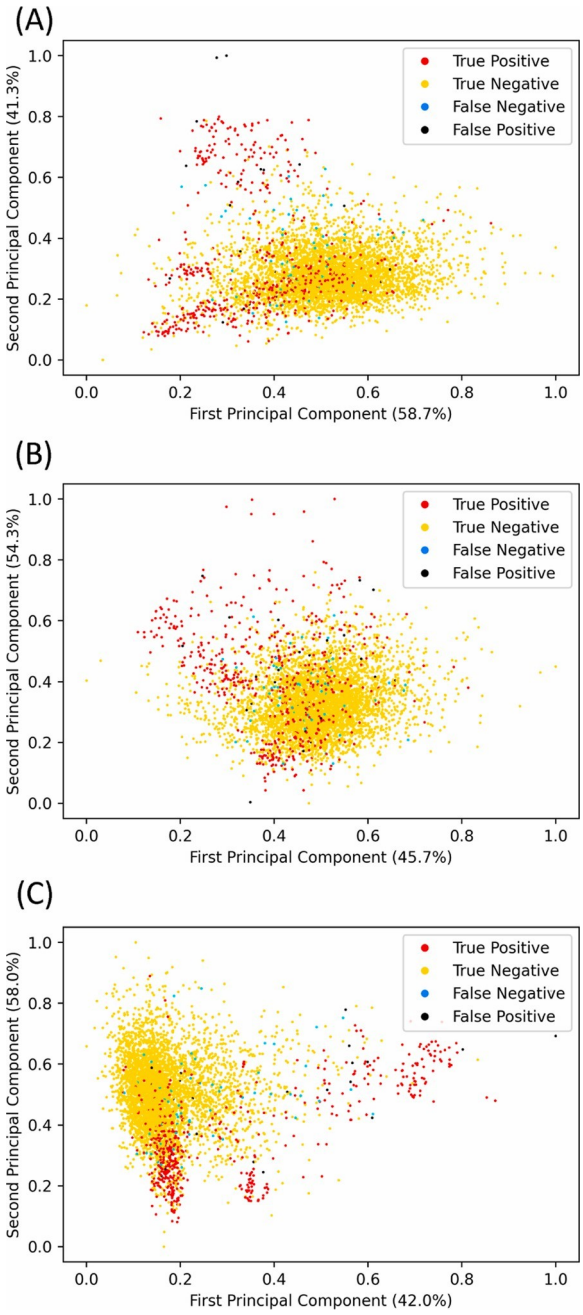

Fig. 3. Principal component analysis of the (A) full input feature set, (B) the subset of only the physicochemical features and (C) the subset of only the compositional descriptors. Anticancer peptides (positives) are coloured red, non-anticancer peptides (negatives) are coloured yellow, false-positives are coloured black, false-negatives are coloured blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4. Anticancer activity prediction

The focus of this novel study was to employ a selection of machine learning methods for the task of predicting peptides' anticancer activity from their primary structure. The classifiers' robustness and predictive power were evaluated by means of 10-fold cross-validation. Additionally, the final ensembles of each cross-validated classifier was evaluated through the use of an external,

independent test set. The accuracy, sensitivity, specificity and Matthews correlation coefficient statistical parameters are reported for each model, along with their respective

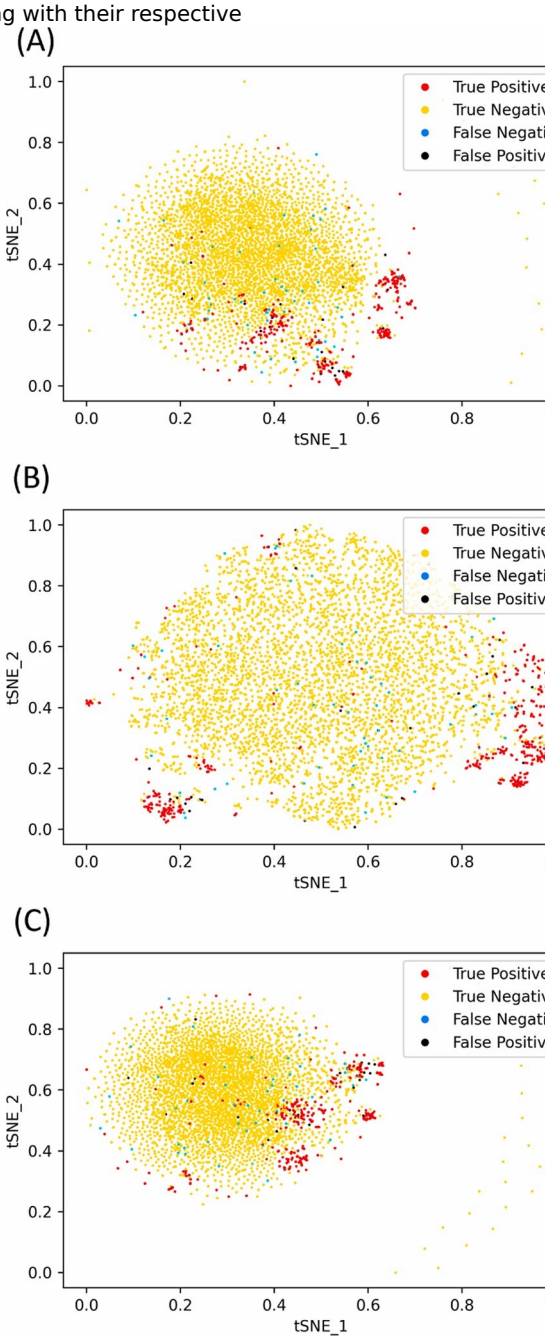

Fig. 4. t-SNE visualization of the (A) full input feature set, (B) the subset of only the physicochemical features and (C) the subset of only the compositional descriptors. Anticancer peptides (positives) are coloured red, non-anticancer peptides (negatives) are coloured yellow, false-positives are coloured black, false-negatives are coloured blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

confidence intervals. A receiver operating characteristic curve (ROC) with a calculated area under the curve (AUC) is also reported for the final neural network model.

The statistical parameters achieved by the support vector machine (SVM), random forest (RF) and neural network (NN) models are detailed in Table 1.

The hyperparameters for the SVM and RF classifiers were optimized using a grid search.

The SVM classifier achieved its best performance with the regularization parameters $C = 1$ and $C = 10$ for the linear and RBF kernels, respectively, and the kernel coefficient $\gamma = 2 \times 10^{-4}$ for the non-linear kernel. The SVM classifiers, regardless of the kernel employed, achieve the worst performance of the three classifier methods studied, with validation accuracies of 96.3% and 96.7%, and MCCs of 0.81 and 0.82, respectively.

The RF classifier performed best, with an accuracy and MCC of 97.3% and 0.86, respectively, when the number of estimators was set to 124, the tree depth was unrestricted, and the maximum number of features set to 50.

The best validation statistics, however, were achieved using the neural network approach, with an accuracy and MCC score of 98.3% and 0.91, respectively. Additionally, the neural network retains the best balance between specificity and sensitivity, making it the most capable predictor. The neural network's predictive power was evaluated further by producing a receiver operating characteristic (ROC) curve, and calculating its associated area under the curve (AUC) (Fig. 5), which corresponds to the likelihood that the neural network will rank a randomly selected anticancer peptide higher than a non-anticancer one. The classifier achieves excellent performance both in cross-validation and external validation, with AUCs of 0.95 and 0.97, respectively.

As the neural network outperformed the SVM and RF approaches, it was selected as the best model for anticancer activity prediction and studied further.

3.5. Descriptor-set specific results

The neural network was trained on feature vectors composed of physicochemical and compositional descriptors. A number of neural networks were trained and evaluated on subsets of this input space, to ascertain what the predictive capacity of each of these subsets is. The validation results of these neural networks, trained on the peptides' physicochemical features, dipeptide composition, tripeptide composition, and dipeptide *g*-gap composition, are detailed in Table 2.

Interestingly, of the reduced subset approaches evaluated, none achieve performance greater than the model trained on all the available information. This finding justifies the prioritization of the hybrid approach, as each of the feature subsets provides the hybrid model with unique information that is required for maximal performance.

3.5.1. Dipeptide and tripeptide composition

Dipeptide and tripeptide compositions are defined as the proportion of a given dipeptide or tripeptide in the sequence. These features are particularly useful as they retain local sequence order information, while also capturing the chemical nature of the peptide. Both models achieve respectable results, with the dipeptide model yielding an accuracy and MCC of 95.56% and 0.77,

respectively, while the tripeptide model yields an accuracy and MCC of 93.97% and 0.70, respectively.

3.5.2. g-Gap composition *g*-gap dipeptide composition can be defined as the proportion of a pair of amino acids separated by 1, 2 or 3 residues. *g*-gap amino acid pairs correspond to residues which may be proximate to one another in three-dimensional space, which is pertinent information for peptides which typically possess some defined secondary structural elements where non-adjoining residues may be connected by hydrogen bonds. Interestingly, training on *g*-gap dipeptide composition results in higher prediction accuracy than training on the conventional dipeptide composition, with an accuracy and MCC of 96.74% and 0.83 respectively. The improvement in classification capability can be attributed to the *g*-gap features being better suited to describing the chemical environment which the structured peptide exposes to the target membrane upon contact, such as a pair of amino acids which are spatially proximate despite being separated by one turn of the α-helix, which has a turn of 3.6 residues.

3.5.3. Physicochemical

The model trained on physicochemical features exhibits high

Table 1
Cross-validation and external validation statistical parameters for the SVM, RF and NN models trained on the ENNAACT and ENNAACT-HARD datasets. Additionally, the ENNAACT NN model is evaluated using the external dataset based on the works of Agrawal et al. [23] and Basith et al. [54].

Model dataset	Method	Acc	Sn	Sp	MCC
Cross-validation					
ENNAACT					
	SVM (linear)	96.31 ±0.52	82.52 ±3.18	98.03 ±0.41	0.81 ±0.02
	SVM (RBF)	96.71 ±0.50	73.04 ±3.71	99.66 ±0.17	0.82 ±0.02
	RF	97.34 ±0.45	79.24 ±3.39	99.59 ±0.19	0.86 ±0.01
	NN	98.29 ±0.36	89.80 ±2.53	99.34 ±0.24	0.91 ±0.01
ENNAACT-HARD	SVM (linear)	87.25 ±1.97	88.16 ±2.70	86.40 ±2.86	0.75 ±0.04
	SVM (RBF)	78.69 ±2.42	95.63 ±1.71	61.90 ±4.05	0.61 ±0.05
	RF	88.60 ±1.88	89.98 ±2.51	87.28 ±2.77	0.77 ±0.04
	NN	93.67 ±1.43	91.06 ±2.39	96.16 ±1.60	0.88 ±0.03
	NN				
ENNAACT+A1		98.15 ±0.35	87.69 ±2.69	99.32 ±0.22	0.90 ±0.01
ENNAACT+A2	NN	93.78 ±1.32	90.92 ±2.11	97.37 ±1.31	0.88 ±0.03
ENNAACT+B	NN	98.04 ±0.42	91.11 ±2.48	99.01 ±0.32	0.91 ±0.01
External validation					
ENNAACT					
	SVM (linear)	97.30 ±1.01	87.02 ±6.28	98.58 ±0.78	0.86 ±0.03
	SVM (RBF)	97.00 ±1.06	76.36 ±7.94	99.57 ±0.43	0.84 ±0.03
	RF	97.48 ±0.97	81.40 ±7.27	99.49 ±0.47	0.87 ±0.03
	NN	97.99 ±0.87	89.09 ±5.83	99.10 ±0.62	0.90 ±0.03
ENNAACT-HARD	SVM (linear)	90.82 ±3.85	91.24 ±5.28	90.39 ±5.61	0.82 ±0.08
	SVM (RBF)	81.48 ±5.18	95.45 ±3.89	66.98 ±8.95	0.65 ±0.10
	RF	87.63 ±4.39	88.18 ±6.03	87.05 ±6.39	0.75 ±0.09
	NN	93.52 ±3.28	90.91 ±5.37	96.23 ±3.63	0.87 ±0.07
	NN				
ENNAACT+A1		92.70 ±2.87	87.85 ±4.76	99.25 ±1.45	0.86 ±0.06
ENNAACT+A2	NN	93.33 ±2.75	90.06 ±4.36	97.76 ±2.50	0.87 ±0.05
ENNAACT+B	NN	96.34 ±0.83	80.93 ±5.02	98.44 ±0.58	0.82 ±0.03

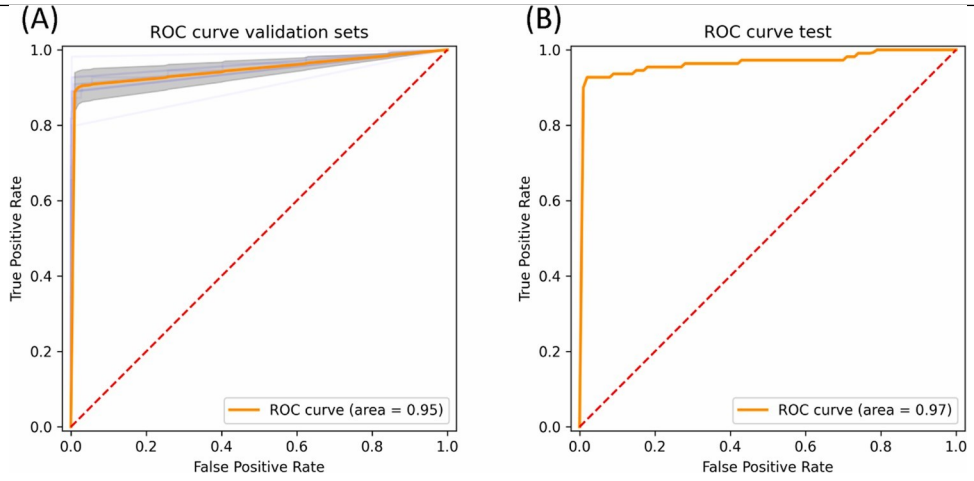

Fig. 5. Receiver operating characteristic (ROC) plots and associated area under the curve (AUC) values for ENNAACT performance on (A) the 10-fold cross-validation set and (B) the external test set.

performance, with an accuracy and MCC of 97.28% and 0.86. This performance is close to that of the hybrid model trained on all available features, highlighting the importance of a peptide possessing the correct physicochemical properties for the exertion of anticancer activity.

4. Discussion

Increasing interest in peptide therapeutics follows an ever- decreasing number of drug approvals and an increasingly large research and development cost base. An ideal peptide anticancer agent should exhibit biological activity against the target cancer cells, while minimizing toxicity towards the non-malignant cells. Peptide toxicity towards eukaryotic cells is often approximated through measuring their hemolytic activity, which the authors of this work have developed a neural network to predict from the primary sequence [47]. Similarly, the focus of this work is on the prediction of peptide anticancer activity from the primary sequence.

A lot of research has been conducted on the structures of natural and engineered bioactive peptides, and the relationship that these structures have to the peptides' biological activities [10,117,118]. Nonetheless, our understanding of the relationship between peptide structure and function remains incomplete. This is especially true in the case of peptide anticancer activity, which compared to peptide antimicrobial activity is poorly understood and less studied, with fewer peptides exhibiting such activity identified in the literature. An *in silico* method for the prediction of peptide anticancer activity on the basis of its primary structure, without the need for expensive chemical synthesis and laboratory evaluation, would provide for a rapid exploration of the peptide chemical space, and a significant cost-saving in the research and development of novel anticancer agents. Machine learning classification of a peptide's anticancer activity is complicated by the comparatively limited amount of available data on anticancer peptide sequences. Unfortunately, testing peptides for anticancer activity is not commonplace, and so the positive dataset employed in this study is limited to 659 peptide sequences with experimentally validated anticancer activity. Furthermore, it is even less commonplace for researchers to report negative anticancer activity results in their studies, which in turn necessitates the use of peptide sequences randomly extracted from UniProt as the negative dataset. While it may initially seem attractive to use the available large datasets of antimicrobial peptides as the negative dataset, it is probable that a significant proportion of those peptides possess anticancer activity that simply has not been tested for in the laboratory.

The aim of the present study was to create a computational method for classifying peptides as anticancer or non-anticancer based on their primary structure, with a view to accelerating the screening for new anticancer agents. Discriminating between the two classes, however, is a challenging task as it requires the ability to selectively differentiate between desirable activity against human cancer cells and toxic activity against non-malignant cells.

The robustness of a neural network classifier is predicated on correctly defining the classification task, specifically the positive and negative datasets. The positive dataset consists of experimentally validated peptide sequences which exhibit anticancer activity. To ensure maximal predictive performance, the positive dataset's peptides' anticancer activity was verified by checking the literature in each instance; in cases where reference to anticancer activity for a given peptide was not found in the literature were excluded from the dataset. The negative dataset, meanwhile, comprises non-secretory peptides extracted from proteins in UniProt. To prevent overfitting, the dataset was subjected to internal redundancy reduction, so that no pair of sequences has a similarity exceeding 90%. The authors believe that as the ENNAACT dataset consists of 659 peptides which have been confirmed to possess anticancer activity, significantly more than similar existing datasets, the dataset represents a major improvement in a field where large quantities of data allow for improved prediction accuracy. With a view to allowing other researchers to improve upon this work, the ENNAACT dataset is available for download both on the ENNAACT web server's website, and as supplementary information to this article.

Once a reliable dataset was assembled from all available sources and validated, peptide sequences were converted into feature vectors comprising their compositional and physicochemical features, and the capacity of support vector machines, random forests and neural networks to predict the peptides' anticancer activity was evaluated by means of 10-fold cross-validation and external validation on an independent test set. The neural network model was

Table 2
Validation statistics achieved by neural networks trained with just single sets of descriptors. The *g*-gap parameter $g = 1,2,3$.

Dataset	Features	Acc (%)	Sn (%)	Sp (%)	MCC
Cross-validation					
ENNAACT					
	Composition, dipeptide	95.56 ±0.57	78.93 ±3.41	97.63 ±0.49	0.77 ±0.02
	Composition, tripeptide	93.97 ±0.66	75.00 ±3.62	96.33 ±0.55	0.70 ±0.02
	g-gap composition, dipeptide	96.74 ±0.49	83.29 ±3.12	98.41 ±0.37	0.83 ±0.02
	Physicochemical	97.28 ±0.45	88.52 ±2.67	98.37 ±0.37	0.86 ±0.01
External validation					
ENNAACT					
	Composition, dipeptide	96.04 ±1.21	76.36 ±7.94	98.49 ±0.80	0.79 ±0.04
	Composition, tripeptide	94.40 ±1.43	73.64 ±8.23	96.98 ±1.13	0.71 ±0.04
	g-gap composition, dipeptide	96.88 ±1.08	84.24 ±6.81	98.45 ±0.81	0.84 ±0.03
	Physicochemical	97.22 ±1.02	86.97 ±6.29	98.49 ±0.80	0.86 ±0.03

found to be the most capable and was consequently retained, optimized and further studied.

The best performing neural network model possesses a 10-fold cross- validated accuracy, sensitivity and specificity of 98.3%, 89.8% and 99.3%, an MCC of 0.91 and an AUC of 0.95. It is clear from the validation statistics that the model is capable of accurately differentiating between anticancer and non-anticancer peptides. The model's validation statistics are comparable to the best-in-class methods.

To confirm that the neural network model is capable of differentiating between peptides with similar compositions, an additional model was created, and trained on the more challenging and class-balanced ENNAACT-HARD dataset. This is an additional dataset which consists of all of the anticancer peptides of the ENNAACT dataset, and for each of those, the most compositionally similar non-anticancer peptide, with compositional similarity defined by the Euclidean distance between their amino acid composition vectors. Despite the compositional similarity between the two classes, a respectable cross-validated accuracy and MCC of 93.7% and MCC of 0.88 are achieved. Nonetheless, the model's accuracy and sensitivity would benefit, however, from a larger repository of experimentally validated anticancer peptides. As the literature on the subject of anticancer peptides expands, the capacity of neural networks to correctly classify anticancer peptides is expected to expand along with it.

Basith et al. have performed an extensive and thorough comparison of currently available anticancer peptide predictors, which included independently evaluating all the publicly available methods on a new independent dataset [54]. Their evaluation identified mACPpred [35] as possessing the best performance with a balanced accuracy (BACC), MCC, Sn and Sp of 87.6%, 0.64, 84.2% and 91.1%, respectively. To fairly extend their evaluation to include ENNAACT, an additional dataset ENNAACT+B was created, which consists of the independent dataset of Basith et al. as the test set, and the peptides of the ENNAACT dataset, redundancy reduced against the test set and internally redundancy reduced, as the training and internal validation set. The model trained on the ENNAACT+B dataset achieved excellent independent test statistics, with a balanced accuracy (BACC), MCC, Sn and Sp of 89.7%, 0.82, 80.9% and 98.4%. These results demonstrate that ENNAACT outperforms the state-of-the-art on this benchmark dataset, which Basith et al. identified as mACPpred, with a BACC, MCC, Sn and Sp of 87.6%, 0.64, 84.2% and 91.1%.

Since the publication of the review of Basith et al., however, Agrawal et al. published their work describing AntiCP 2.0, which represents the current state-of-the-start, with an Acc, Sn, Sp and MCC of 92.0%, 92.3%, 91.8% and 0.84, respectively [23]. To compare the performance of ENNAACT to AntiCP 2.0, the ENNAACT+A1 dataset was constructed, which uses the ENNAACT sequences for training and validation, and the AntiCP 2.0 test sequences for its independent test set, and a model was trained and evaluated with this dataset. Additionally, for a more thorough comparison, the ENNAACT+A2 model was trained on the AntiCP 2.0 training set, and evaluated with the AntiCP 2.0 test set. Both models surpass the performance achieved by AntiCP 2.0 on all measures except sensitivity, with Acc of 92.7% and 93.3%, and MCC of 0.86 and 0.87, respectively.

Unsupervised dimensionality reduction methods were employed for the visualization of the feature space, both in its entirety and in its compositional and physicochemical subsets. Principal component analysis (PCA) (Fig. 3) and t-Distributed Stochastic Neighbour Embedding (t-SNE) (Fig. 4) were employed, with the latter achieving a more clear, albeit still incomplete, separation between the positive and negative classes. The physicochemical subset's t-SNE plot specifically is the most clearly separated. To visualize the misclassified peptides, the false positives and negatives were highlighted on the PCA and t-SNE plots, making clear that prediction errors are largely due to peptides' physicochemical and compositional similarity to peptides more typical of the opposite class.

To ascertain their respective predictive powers, the neural network model was trained on reduced feature spaces. Judging by 10-fold cross- validation accuracies, the model trained on physicochemical descriptors is the most powerful, although the model trained on *g-gap* dipeptide composition is only slightly less accurate. These results highlight the relationship between the peptides' physicochemical properties and three-dimensional structure, and the anticancer potency.

To conclude, the limited quantity of available experimentally validated data and the incomplete understanding of the mechanism of peptide anticancer activity continue to pose challenges for the research community. In an effort to overcome these challenges, this study presented ENNAACT, a novel *in silico* peptide anticancer activity classifier.

The classifier, which employs a deep neural network architecture and benefits from a newly constructed dataset with improved data quality and quantity projected on a rich feature-space, achieves predictive power on-par with the state-of-the-art. ENNAACT is available online for use by academic researchers at https://research.timmons.eu/ennaact. The authors hope that the this work will facilitate better *in-silico* design of novel peptide-based anticancer therapeutics, thereby enabling an acceleration of the design phase and a reduction in its cost, helping to drive medicinal chemistry into an unprecedented revolution.

5. Web server implementation

As a user-friendly web server is essential for realizing the impact of bioinformatic research, the classifier is available as an easy to use, publicly accessible web server online at https://research.timmons.eu/e nnaact, for the benefit of the scientific community. The web server is capable of predicting peptides' anticancer activity based on the primary sequence, N-terminal acetylation and C-terminal amidation modifications. Input peptide sequences are restricted to the 20 natural amino acids; support for non-natural amino acids is not available. Many features are included in the web server, and models trained on the redundancy reduced ENNAACT dataset, the ENNAACT-HARD dataset, and the benchmarking

ENNAACT+A1, ENNAACT+A2 and ENNAACT+B datasets are available for prediction. *5.1. Anticancer activity prediction*

Anticancer activities can be predicted for both individual and multiple sequences. The user should provide the peptide sequences input in FASTA format and choose the model for prediction, and the probability of each sequence possessing anticancer activity will be returned from the neural network's prediction. Probabilities are given on a scale of 0-1, whereby 0 and 1 are most probably non-anticancer, and most probably anticancer, respectively. Batch submission of sequences in FASTA format is possible. The maximum batch size is variable depending on the length of the sequences; longer sequences necessitate smaller batch sizes.

5.2. Mutation analysis

Individual peptide sequences, input in FASTA format, may be subjected to a mutation analysis. To carry out a mutation analysis, the user should select the mutation analysis option, input the residue number chosen for mutation, and run the prediction. Mutant peptide sequences will be generated by substituting the residue at the given position with each of the remaining 20 natural amino acids. The probability of each of the peptide's mutants possessing anticancer activity will be returned by the chosen neural network.

5.3. Residue scan

Residue scans, such as an alanine scan, are available for individual FASTA format sequences. Users should select the residue scan option, choose the amino acid residue they wish to scan with, and run the prediction. The probability of anticancer activity for the native sequence, and its mutants attained by substituting successive residues with the selected amino acid residue, will be returned by the chosen neural network.

Data availability

All data generated or analysed during this study are available for download at https://research.timmons.eu/ennaact.

Conflict of interest

The authors have no competing interests to declare.

Acknowledgements

The authors would also like to thank University College Dublin for the Research Scholarship granted to P.B.T.

References

[1] A. Jemal, et al., Global cancer statistics, CA: Cancer J. Clin. 61 (2011) 69–90.
[2] J. Ferlay, et al., Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008, Int. J. Cancer 127 (2010) 2893–2917.
[3] C. Holohan, S. Van Schaeybroeck, D.B. Longley, P.G. Johnston, Cancer Drug Resistance: An Evolving Paradigm, 2013.
[4] Y.-B. Huang, L.-Y. He, H.-Y. Jiang, Y.-X. Chen, Role of helicity on the anticancer mechanism of action of cationic-helical peptides, Int. J. Mol. Sci. 13 (2012) 6849–6862.
[5] L. Steinstraesser, U. Kraneburg, F. Jacobsen, S. Al-Benna, Host Defense Peptides and Their Antimicrobial-Immunomodulatory Duality, 2011.
[6] M. Mahlapuu, J. Håkansson, L. Ringstad, C. Bjorn, Antimicrobial Peptides: An¨ Emerging Category of Therapeutic Agents, 2016.
[7] R.E. Hancock, G. Diamond, The role of cationic antimicrobial peptides in innate host defences, Trends Microbiol. 8 (2000) 402–410.
[8] L. Otvos, Peptide-Based Drug Design: Here and Now, 2008.
[9] N. Papo, Y. Shai, Host Defense Peptides As New Weapons in Cancer Treatment, 2005.
[10] S. Benetti, P.B. Timmons, C.M. Hewage, NMR model structure of the antimicrobial peptide maximin 3, Eur. Biophys. J. 48 (2019) 203–212.
[11] D. Andreu, L. Rivas, Animal antimicrobial peptides: an overview, Biopolymers 47 (1998) 415–433.
[12] 'T. Van, W. Hof, E.C. Veerman, E.J. Heimerhorst, A.V. Nieuw Amerongen, Antimicrobial Peptides: Properties and Applicability, 2001.
[13] D.W. Hoskin, A. Ramamoorthy, Studies on Anticancer Activities of Antimicrobial Peptides, 2008.
[14] J.S. Mader, D.W. Hoskin, Cationic Antimicrobial Peptides As Novel Cytotoxic Agents for Cancer Treatment, 2006.
[15] S.R. Dennison, F. Harris, D.A. Phoenix, Investigations into the potential anticancer activity of Maximin H5, Biochimie 137 (2017) 29–34.
[16] H.M. Ellerby, et al., Anti-cancer activity of targeted pro-apoptotic peptides, Nat. Med. 5 (1999) 1032–1038.
[17] F. Harris, S.R. Dennison, J. Singh, D.A. Phoenix, On the selectivity and efficacy of defense peptides with respect to cancer cells, Med. Res. Rev. 33 (2013) 190–234.
[18] F. Schweizer, Cationic Amphiphilic Peptides With Cancer-Selective Toxicity, 2009.
[19] D. Gaspar, A. Salom´e Veiga, M.A. Castanho, From Antimicrobial to Anticancer Peptides. A Review, 2013.
[20] K. Kozłowska, J. Nowak, B. Kwiatkowski, M. Cichorek, ESR study of plasmatic membrane of the transplantable melanoma cells in relation to their biological properties, Exp. Toxicol. Pathol. 51 (1999) 89–92.
[21] M. Sok, M. Sentjurc, M. Schara, Membrane fluidity characteristics of human lung cancer, Cancer Lett. 139 (1999) 215–220.
[22] A. Tyagi, et al., In silico models for designing and discovering novel anticancer peptides, Sci. Rep. 3 (2013) 1–8.
[23] P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, G.P.S. Raghava, AntiCP 2.0: an updated model for predicting anticancer peptides, Brief. Bioinformatics (2020).
[24] W. Chen, H. Ding, P. Feng, H. Lin, K.C. Chou, iACP: a sequence-based tool for identifying anticancer peptides, Oncotarget 7 (2016) 16895–16909.
[25] V. Saravanan, P.T. Lakshmi, ACPP: A web server for prediction and design of anti- cancer peptides, Int. J. Pept. Res. Therapeut. 21 (2015) 99–106.
[26] S. Akbar, M. Hayat, M. Iqbal, M.A. Jan, iACP-GAEnsC: evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space, Artif. Intell. Med. 79 (2017) 62–70.
[27] B. Manavalan, et al., MLACP: Machine-learning-based prediction of anticancer peptides, Oncotarget 8 (2017) 77121–77136.
[28] L. Xu, G. Liang, L. Wang, C. Liao, A novel hybrid sequence-based model for identifying anticancer peptides, Genes 9 (2018).
[29] M. Kabir, et al., Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information, Chemometr. Intell. Lab. Syst. 182 (2018) 158–165.
[30] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, ACPred: a computational tool for the prediction and analysis of anticancer peptides, Molecules 24 (2019).
[31] B. Rao, C. Zhou, G. Zhang, R. Su, L. Wei, ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides, Brief. Bioinformatics (2019).

[32] L. Wei, C. Zhou, H. Chen, J. Song, R. Su, ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides, Bioinformatics (Oxford, England) 34 (2018) 4007–4016.
[33] C. Wu, R. Gao, Y. Zhang, Y. De Marinis, PTPD: Predicting therapeutic peptides by deep learning and word2vec, BMC Bioinformatics 20 (2019) 456.
[34] L. Wei, C. Zhou, R. Su, Q. Zou, J. Hancock, PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning, Bioinformatics 35 (2019) 4272–4280.
[35] V. Boopathi, et al., mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides, Int. J. Mol. Sci. 20 (2019).
[36] H.C. Yi, et al., ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation, Mol. Ther. Nucleic Acids 17 (2019) 1–9.
[37] F.M. Li, X.Q. Wang, Identifying anticancer peptides by using improved hybrid compositions, Sci. Rep. 6 (2016) 1–6.
[38] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, J. Theoret. Biol. 341 (2014) 34–40.
[39] F. Khan, S. Akbar, A. Basit, I. Khan, H. Akhlaq, Identification of anticancer peptides using optimal feature space of Chou's split amino acid composition and support vector machine. ACM International Conference Proceeding Series, 91–96 (Association for Computing Machinery, New York, New York, USA) (2017).
[40] G. Gabernet, et al., In silico design and optimization of selective membranolytic anticancer peptides, Sci. Rep. 9 (2019) 11282.
[41] Q. Li, W. Zhou, D. Wang, S. Wang, Q. Li, Prediction of anticancer peptides using a low-dimensional feature model, Front. Bioeng. Biotechnol. 8 (2020) 892.
[42] W. Shoombuatong, N. Schaduangrat, C. Nantasenamat, Unraveling the bioactivity of anticancer peptides as deduced from machine learning, EXCLI J. 17 (2018) 734–752.
[43] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins Struct. Funct. Genet. 43 (2001) 246–255.
[44] M. Zeng, et al., Protein-protein interaction site prediction through combining local and global features with deep neural networks, Bioinformatics (2019).
[45] M. Oti, S. Ballouz, M.a. Wouters, In silico tools for gene discovery, Methods Mol. Biol. 760 (2011) 175–187.
[46] T.A. Holton, G. Pollastri, D.C. Shields, C. Mooney, CPPpred: prediction of cell penetrating peptides, Bioinformatics 29 (2013) 3094–3096.
[47] P.B. Timmons, C.M. Hewage, HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks, Sci. Rep. 10 (2020) 10869.
[48] K.C. Chou, Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition, 2011.
[49] M. Pirtskhalava, et al., Erratum: DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides (Nucleic Acids Research 44 (D1104-D1112) DOI: 10.1093/nar/gkv1174), Nucleic Acids Res. 44 (2016) 6503.
[50] J.-H. Jhong, et al., dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data, Nucleic Acids Res. 47 (2019) D285–D297.
[51] A. Tyagi, et al., CancerPPD: a database of anticancer peptides and proteins, Nucleic Acids Res. 43 (2015) D837–D843.
[52] Q. Li, et al., BioPepDB: an integrated data platform for food-derived bioactive peptides, Int. J. Food Sci. Nutr. 69 (2018) 963–968.
[53] R. Bals, Epithelial Antimicrobial Peptides in Host Defense Against Infection, 2000.
[54] S. Basith, B. Manavalan, T. Hwan Shin, G. Lee, Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening, Med. Res. Rev. 40 (2020) 1276–1314.
[55] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[56] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics 26 (2010) 680–682.
[57] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next- generation sequencing data, Bioinformatics 28 (2012) 3150–3152.
[58] M. Kumar, V. Thakur, G.P. Raghava, COPid: composition based protein identification, In Silico Biol. 8 (2008) 121–128.
[59] P. Agrawal, et al., In silico approach for prediction of antifungal peptides, Front. Microbiol. 9 (2018) 323.
[60] P. Agrawal, S. Kumar, A. Singh, G.P. Raghava, I.K. Singh, NeuroPIpred: a tool to predict, design and scan insect neuropeptides, Sci. Rep. 9 (2019) 5129.
[61] K.K. Dey, D. Xie, M. Stephens, A new sequence logo plot to highlight enrichment and depletion, BMC Bioinformatics 19 (2018) 473.
[62] A.T. Müller, G. Gabernet, J.A. Hiss, G. Schneider, modlAMP: Python for antimicrobial peptides, Bioinformatics (Oxford, England) 33 (2017) 2753–2755.
[63] A. Ikai, Thermostability and aliphatic index of globular proteins, J. Biochem. 88 (1980) 1895–1898.
[64] J.R. Lobry, C. Gautier, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, Nucleic Acids Res. 22 (1994) 3174–3180.
[65] H.G. Boman, D. Wade, I.A. Boman, B. Wåhlin, R.B. Merrifield, Antibacterial and antimalarial properties of peptides that are cecropin-melittin hybrids, FEBS Lett. 259 (1989) 103–106.
[66] P. Argos, J.K. Rao, P.A. Hargrave, Structural prediction of membrane-bound proteins, Eur. J. Biochem. 128 (1982) 565–575.
[67] D. Eisenberg, R.M. Weiss, T.C. Terwilliger, W. Wilcox, Hydrophobic moments and protein structure, Farad. Symp. Chem. Soc. 17 (1982) 109–120.
[68] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[69] T.P. Hopp, K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences, Proc. Natl. Acad. Sci. U.S.A. 78 (1981) 3824–3828.
[70] J.L. Cornette, et al., Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins, J. Mol. Biol. 195 (1987) 659–685.
[71] J.M. Zimmerman, N. Eliezer, R. Simha, The characterization of amino acid sequences in proteins by statistical methods, J. Theoret. Biol. 21 (1968) 170–201.
[72] T.L. McMeekin, M. Wilensky, M.L. Groves, Refractive indices of proteins in relation to amino acid composition and specific volume, Biochem. Biophys. Res. Commun. 7 (1962) 151–156.
[73] R. Bhaskaran, P.K. Ponnuswamy, Positional flexibilities of amino acid residues in globular proteins, Int. J. Pept. Protein Res. 32 (1988) 241–255.
[74] M. Levitt, M. Levitt, conformational preferences of amino acids in globular proteins, Biochemistry 17 (1978) 4277–4285.
[75] G. Zhao, E. London, An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity, Protein Sci. 15 (2006) 1987–2001.
[76] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862–864.
[77] D. Jureti´c, D. Vuki˘cevi´c, N. Ili´c, N. Antcheva, A. Tossi, Computational design of highly selective antimicrobial peptides, J. Chem. Inform. Model. 49 (2009) 2873–2882.
[78] A. Senes, et al., Ez, a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices, J. Mol. Biol. 366 (2007) 436–448.
[79] E.R. Collantes, W.J. Dunn, Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues, J. Med. Chem. 38 (1995) 2705–2713.
[80] C. Raychaudhury, A. Banerjee, P. Bag, S. Roy, Topological shape and size of peptides: identification of potential allele specific helper T cell antigenic sites, J. Chem. Inform. Comput. Sci. 39 (1999) 248–254.

[81] A. Zaliani, E. Gancia, MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies, J. Chem. Inform. Comput. Sci. 39 (1999) 525–533.
[82] C.P. Koch, et al., Scrutinizing MHC-I binding peptides and their limits of variation, PLoS Comput. Biol. 9 (2013) e1003088.
[83] M. Cocchi, E. Johansson, Amino acids characterization by grid and multivariate data analysis, Quant. Struct.-Activ. Relation. 12 (1993) 1–8.
[84] S. Hellberg, M. Sjost¨ rom, B. Skagerberg, S. Wold, Peptide quantitative structure-¨ activity relationships, a multivariate approach, J. Med. Chem. 30 (1987) 1126–1135.
[85] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjost¨ rom, S. Wold, New chemical¨ descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids, J. Med. Chem. 41 (1998) 2481–2491.
[86] S. Kawashima, et al., AAindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2008) D202–D205.
[87] L.A. Kuhn, C.A. Swanson, M.E. Pique, J.A. Tainer, E.D. Getzoff, Atomic and residue hydrophilicity in the context of folded protein structures, Proteins: Struct. Funct. Bioinformatics 23 (1995) 536–547.
[88] P. Klein, M. Kanehisa, C. DeLisi, Prediction of protein function from sequence properties, Discriminant analysis of a data base, Biochim. Biophys. Acta (BBA)/ Protein Struct. Mol. 787 (1984) 221–226.
[89] C.R. Woese, Evolution of the genetic code, Die Naturwissensch. 60 (1973) 447–459.
[90] W.R. Krigbaum, A. Komoriya, Local interactions as a structure determinant for protein molecules: II, BBA – Protein Struct. 576 (1979) 204–228.
[91] J.-L. Fauchere, V. Pliska, Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides, Eur. J. Med. Chem. 18 (1983) 369–375.
[92] M.C. Wilce, M.I. Aguilar, M.T. Hearn, Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides, Anal. Chem. 67 (1995) 1210–1219.
[93] H. Naderi-Manesh, M. Sadeghi, S. Arab, A.A. Moosavi Movahedi, Prediction of protein surface accessibility with information theory, Proteins: Struct. Funct. Genet. 42 (2001) 452–459.
[94] J.M. Parker, D. Guo, R.S. Hodges, New hydrophilicity scale derived from high- performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites, Biochemistry 25 (1986) 5425–5432.
[95] V. Pli˘ska, M. Schmidt, J.L. Fauch`ere, Partition coefficients of amino acids and hydrophobic parameters π of their side-chains as measured by thin-layer chromatography, J. Chromatogr. A 216 (1981) 79–92.
[96] H.R. Guy, Amino acid side-chain partition energies and distribution of residues in soluble proteins, Biophys. J. 47 (1985) 61–70.
[97] M. Charton, Protein folding and the genetic code: an alternative quantitative model, J. Theoret. Biol. 91 (1981) 115–123.
[98] R. Aurora, G.D. Rose, Helix capping, Protein Sci. 7 (1998) 21–38.
[99] H. Zhou, Y. Zhou, quantifying the effect of burial of amino acid residues on protein stability, Proteins: Struct. Funct. Genet. 54 (2004) 315–322.
[100] M. Oobatake, Y. Kubota, T. Ooi, Optimization of Amino Acid Parameters for Correspondence of Sequence to Tertiary Structures of Proteins, Tech. Rep., 1985.
[101] P.D. Thomas, K.A. Dill, An iterative method for extracting energy-like quantities from protein structures, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 11628–11633.
[102] D. Veltri, U. Kamath, A. Shehu, Deep learning improves antimicrobial peptide recognition, Bioinformatics 34 (2018) 2740–2747.
[103] J. Shen, et al., Predicting protein–protein interactions based only on sequences information, Proc. Natl. Acad. Sci. U.S.A. 104 (2007) 4337–4341.
[104] H. Ding, P.M. Feng, W. Chen, H. Lin, Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis, Mol. Biosyst. 10 (2014) 2229–2235.
[105] J. Dong, et al., PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions, J. Cheminformatics 10 (2018) 16.
[106] D.S. Cao, et al., PyDPI: freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies, J. Chem. Inform. Model. 53 (2013) 3086–3096.
[107] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (1901) 559–572.
[108] L. Van Der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2625.
[109] C. Cortes, Support-Vector Networks. Tech. Rep., 1995.
[110] T.K. Ho, Random decision forests. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, vol. 1 of ICDAR'95, IEEE Computer Society, Washington, DC, USA, 1995, pp. 278–282.
[111] B.W. White, F. Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, vol. 76, Spartan Books, New York, 1963.
[112] M. Abadi, et al., TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2016 eprint1603.04467.
[113] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, 32nd International Conference on Machine Learning, ICML 2015, vol. 1 (2015) 448–456, eprint1502.03167.
[114] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[115] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, 3rd International Conference on Learning Representations, ICLR 2015 – Conference Track Proceedings (2015) eprint1412.6980.
[116] K.C. Chou, Low-Frequency Collective Motion in Biomacromolecules and Its Biological Functions, 1988.
[117] P.B. Timmons, D. O'Flynn, J.M. Conlon, C.M. Hewage, Structural and positional studies of the antimicrobial peptide brevinin-1BYa in membrane-mimetic environments, J. Peptide Sci. 25 (2019) e3208.
[118] P.B. Timmons, D. O'Flynn, J.M. Conlon, C.M. Hewage, Insights into conformation and membrane interactions of the acyclic and dicarba-bridged brevinin-1BYa antimicrobial peptides, Eur. Biophys. J. 48 (2019) 701–710.